Enhancing Deeper Emotional Support Through Multilingual Emotional Validation in Dialogue System

Zi Haur Pang, Yahui Fu, Koji Inoue, Tatsuya Kawahara
Graduate School of Informatics, Kyoto University

Research Background

Emotional support *- Act of providing care and understanding to help someone cope with difficult feelings/situation*

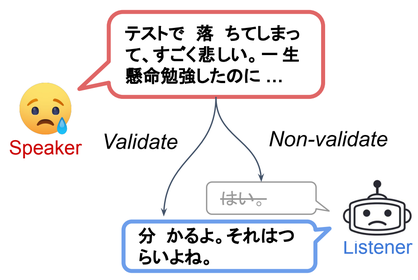
Conventional empathetic responses often fall short for individuals who suppress emotions due to stress or adverse life experiences

⬇

The importance of being understood and accepted is paramount, users want acknowledgment that it's okay to feel their emotions

Emotional Validation

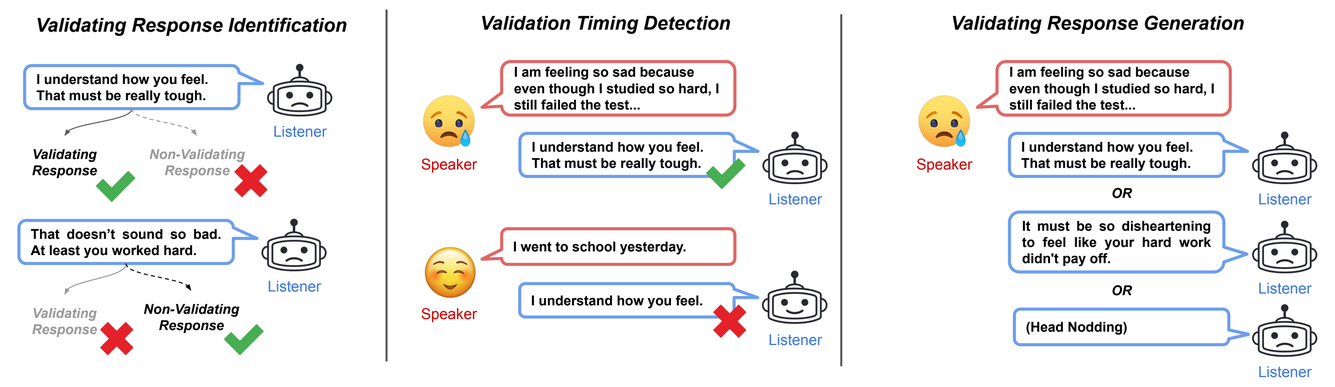
Communication technique in psychotherapy where we recognize, understand, and acknowledge others' emotional states, thoughts, and actions



Research Objective

To investigate **how emotional validation is expressed and timed across languages**, in order to improve **cross-lingual support systems** and reduce inequities caused by language barriers in emotional support communication

Task Description



What is validating response? **When** to express validation? **How** to express validating response?

Dataset Construction

Multilingual - Empathetic Dialogue Emotional Support Conversation (M-EDESConv) *[Text]*

Japanese Empathetic Dialogues (JA) + Empathetic Dialogues (EN)

Emotional Support Conversation (EN) → *GPT-4.1 mini* → Japanese Emotional Support Conversation (JA)

Validation: 46002 (36%)
Non-Validation: 80551 (64%)

Multilingual - TUT Emotional Storytelling Corpus (M-TESC) *[Speech]*

TUT Emotional Storytelling Corpus (JA) → *GPT-4.1 mini* → English TUT Emotional Storytelling Corpus (EN)

Validation: 2052 (34%)
Non-Validation: 2028 (66%)

Objective Evaluation (COMETKiwi):
→ 0.847 for utterance, 0.848 for response

Human Evaluation (Accuracy, Contextual Consistency, Emotional Consistency, Fluency and Readability, Cultural Appropriateness) :
→ 5.63/7 (IAA=0.344, fair agreement)

Validating Response Identification

Manual annotate 3000 utterances (1500 per language) per source (≈ 2% of EmpatheticDialogue and 3% of ESConv)
→ 1884 validation (32%), 3921 non-validation (68%)

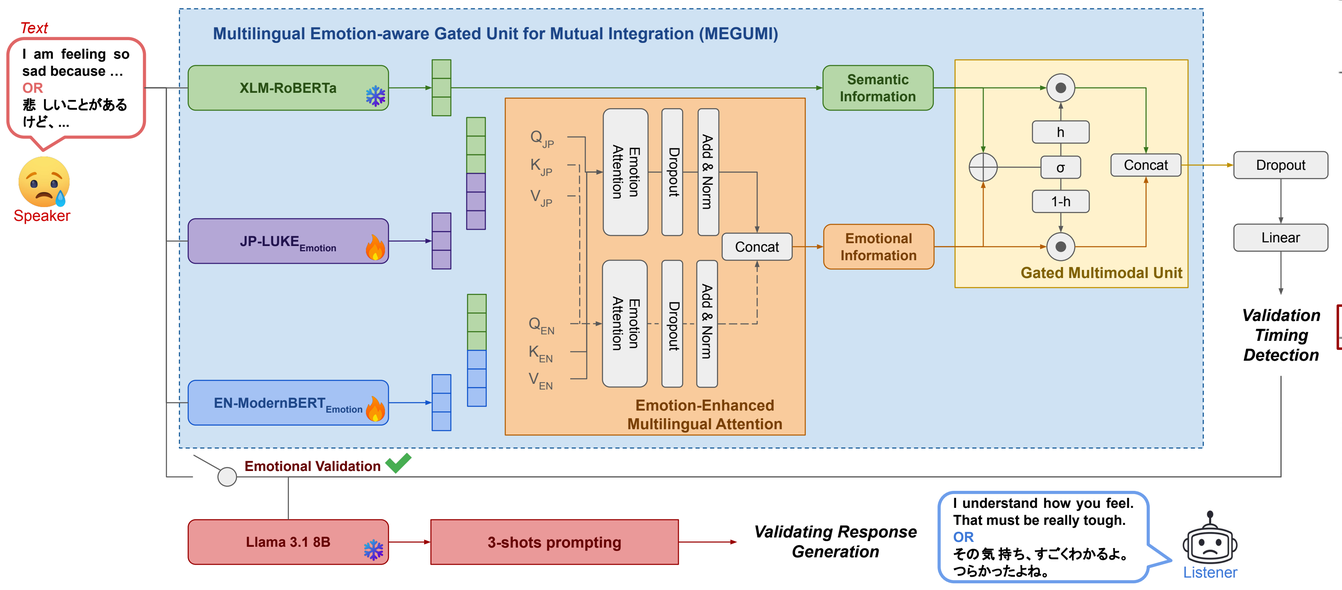
	Macro Average			Target Class		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Random Baseline	50.00	50.00	48.38	32.45	50.13	39.40
mBERT	74.33	74.51	74.30	70.24	76.32	73.15
Llama 3.1 8b						
- *Zero-shot*	54.85	53.39	41.04	34.37	85.61	49.04
- *3-shot*	61.03	52.77	32.18	33.82	**97.88**	50.27
- *LoRA*	58.36	51.81	32.96	37.16	97.02	53.74
GPT 4.1 Nano						
- *Zero-shot*	64.49	65.04	57.98	42.71	85.19	56.89
- *3-shot*	67.21	69.57	66.57	50.36	74.60	60.13
XLM-RoBERTa	**86.42**	**85.30**	**85.28**	**80.66**	93.64	**86.67**

Human evaluation using random 100 samples for each language:
→ 85% Matched (Cohen's κ = 0.752, substantial agreement)

⬇

Utilize **XLM-RoBERTa model** for further automatic annotation

Validating Timing Detection



	M-EDESConv						M-TESC					
	Macro Average			Target Class			Macro Average			Target Class		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
Random Baseline	50.20	50.21	49.23	36.45	50.35	42.30	50.26	50.29	49.03	34.41	50.02	40.77
mBERT	62.07	62.63	59.10	45.16	74.15	56.14	53.28	53.39	53.29	38.29	41.35	39.76
XLM-RoBERTa	62.78	63.13	59.03	45.19	76.96	56.94	55.29	55.74	55.10	40.24	50.00	44.60
Llama 3.1 8b												
- *Zero-shot*	57.33	52.81	37.18	37.72	93.75	53.79	56.84	54.04	40.76	36.31	89.58	51.68
- *3-shot*	57.36	52.40	35.69	37.49	**94.84**	53.73	55.93	51.38	31.53	34.81	**96.29**	51.13
- *LoRA*	50.67	50.25	34.42	36.40	90.81	51.97	50.92	51.02	47.99	33.83	58.44	42.86
GPT 4.1 Nano												
- *Zero-shot*	58.42	57.74	51.68	41.43	79.25	54.41	56.71	57.31	54.04	39.95	66.46	49.90
- *3-shot*	58.87	56.39	46.75	40.02	87.04	54.99	**58.34**	**57.65**	50.14	39.01	80.93	**52.65**
MEGUMI (Ours)	**63.94**	**65.02**	**63.71**	**51.07**	66.11	**57.62**	56.86	57.36	**56.89**	**41.44**	48.70	44.78

Main result (Underline = statistically significant)

	Criteria			Macro Average			Target Class		
	EE	EEMA	GMU	Precision	Recall	F1-Score	Precision	Recall	F1-Score
XLM-RoBERTa	-	x	x	56.78	55.43	47.29	39.60	82.42	53.50
+ Mono-EN	EN	x	x	57.21	57.40	57.27	45.10	48.23	46.61
+ Mono-JP	JP	x	x	57.99	58.63	56.86	43.97	62.88	51.75
+ Multi-Concat	Both	x	x	62.73	63.34	59.80	46.75	73.86	57.26
+ Multi-EEMA	Both	v	x	62.83	63.85	62.48	49.70	65.31	56.45
MEGUMI (Ours)	Both	v	v	63.94	65.02	63.71	51.07	66.11	57.62

Ablation study

Human evaluation using random 100 samples for each language:
→ 72% Matched (Cohen's κ = 0.490, fair agreement)

Validating Response Generation

EmoValidBench - evaluate LLM's ability in validating response generation

	Semantics			Diversity		Empathy		
	$BERT_{Pre}$	$BERT_{Rec}$	$BERT_{F1}$	D1	D2	ER	IP	EX
Llama 3.1 8b								
- *Zero-shot*	87.82	89.60	88.67	4.88	23.96	52.90	70.94	67.14
- *3-shot*	89.11	89.03	89.03	5.52	25.97	54.66	71.54	69.59
- *LoRA*	89.32	89.30	89.29	12.05	30.60	59.52	65.36	64.43
GPT 4.1 Nano								
- *Zero-shot*	88.06	89.75	88.87	4.80	22.37	51.95	73.17	69.97
- *3-shot*	88.61	89.97	89.26	4.58	21.42	52.32	72.29	68.76
- *CoT*	87.15	89.19	88.13	4.83	24.53	50.72	73.30	71.21

Human Evaluation (Naturalness, Contextual Understanding, Emotional Understanding, Non-judgemental Tone, Mindful Presence, Perceived Validation) :
→ Ground Truth 5.77/7 (IAA = 0.41), LLama 5.21/7 (IAA = 0.45), GPT 5.57/7 (IAA = 0.32)
→ Comparable: Naturalness, Contextual Understanding, Non-judgemental Tone
→ Room for Improvement: Emotional Understanding, Mindful Presence, Perceived Validation

Case Study, Conclusion & Future Work

Model	Text
Utterance	Yeah they must practice a lot. I would be afraid of getting trampled
Ground Truth	Me too, that would be terrible.
Llama 3.1 8B	I can understand that. I would feel the same way.
GPT-4.1 Nano	That makes sense, it's understandable to feel worried about that.

→ Formalize **Emotional Validation** task in dialogue system
→ Construct **M-EDESConv** & **M-TESC** dataset
→ Propose **MEGUMI for validation timing detection** in multilingual settings
→ Propose **EmoValidBench** to evaluate LLM's ability in such domain
→ Implement the model in **conversational robots for real-life interaction** to reduce emotional burdens of the user